Data Assimilation Research Centre

Nonlinear data assimilation for ocean applications

Peter Jan van Leeuwen, Manuel Pulido, Jacob Skauvold, Javier Amezcua, Polly Smith, Met Ades, Mengbin Zhu

Colorado State University

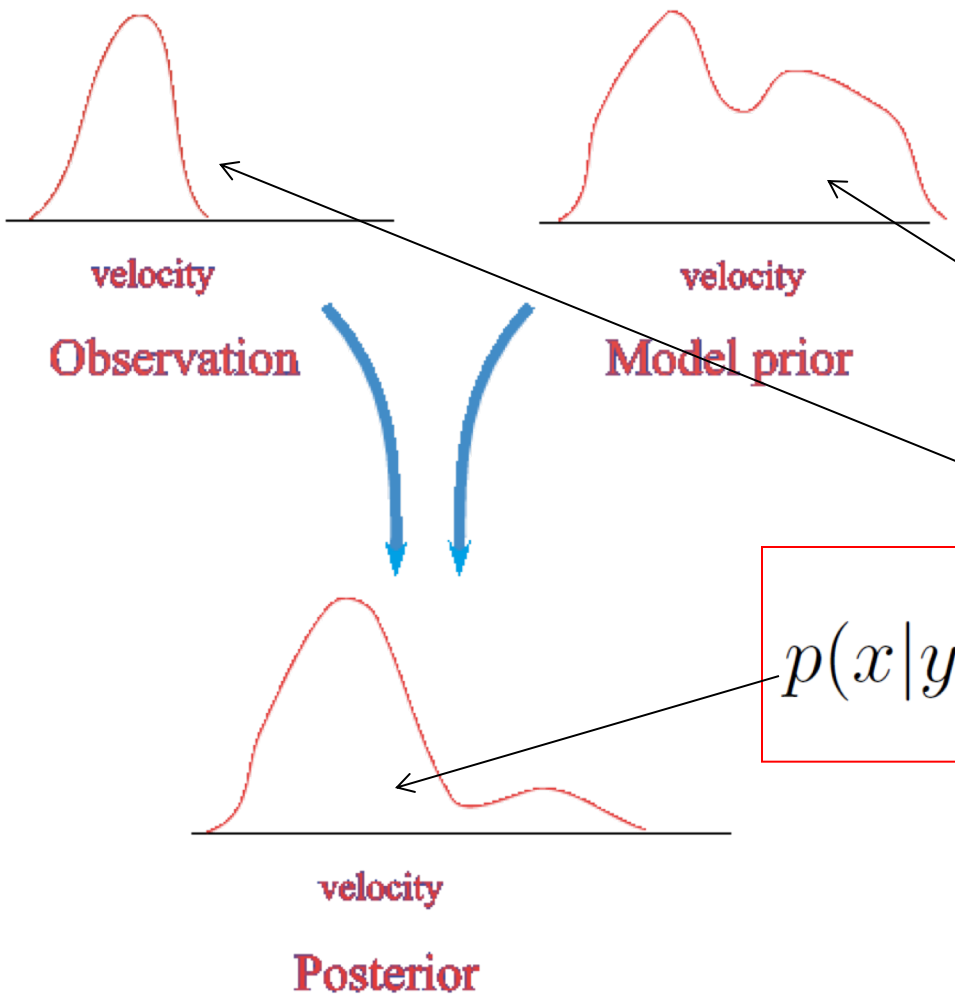
erc
European Research Council
Established by the European Commission

University of Reading

Comparing observations and models

- Ensure they represent the same thing
- Comparison is useless unless uncertainties in observations and models are taken into account (representation errors)
- Data assimilation does that in a systematic way
- DA can be used for:
  - Forecasting
  - Study what real ocean has done over time period
  - Model parameter estimation
  - Systematic model improvement

What is data assimilation?



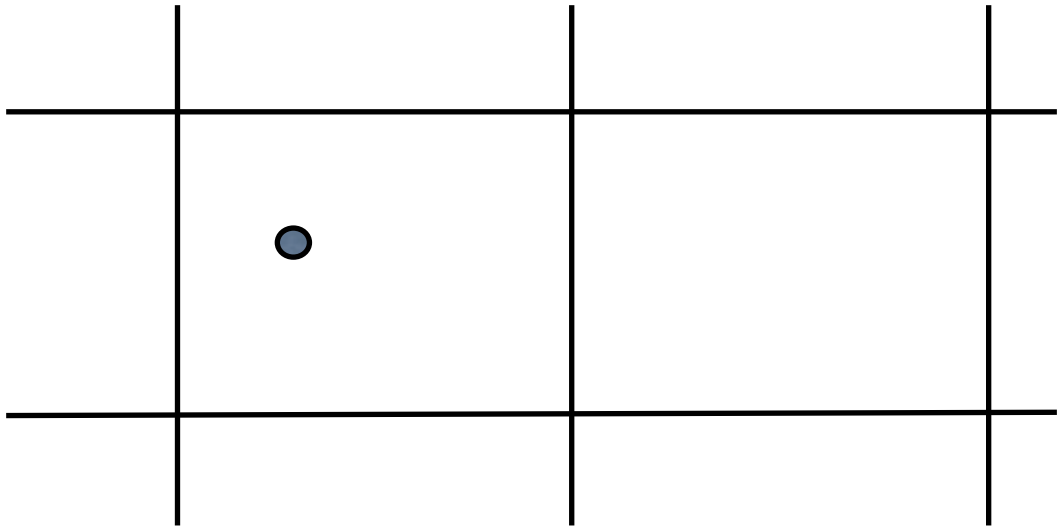
Bayes theorem:

$$p(x|y) = \frac{p(y|x)p(x)}{\int p(y|x)p(x)\ dx}$$

The solution is a pdf!

What are representation errors?

Extra uncertainty that arises in data assimilation because model and observations have different representation of reality.



How do they arise in DA?

Start from Bayes Theorem:

$$\boxed{p(x|y) = \frac{p(y|x)p(x)}{p(y)}}$$

Representation differences between model and observations, so representation errors, arise in the likelihood, not in the prior!

What does the likelihood mean?

Observation equation:

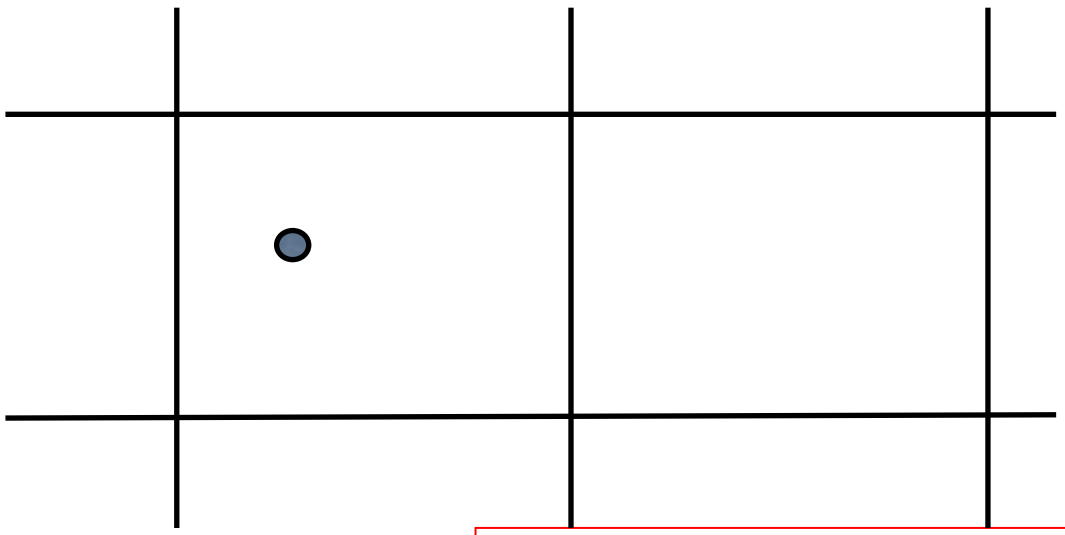
$$\boxed{y = H(x_t) + \epsilon}$$

However, the observations do appear as $p(y|x)$ in Bayes Theorem, with each possible model state *x* as given, not the truth x_t.

Interpretation: $p(y|x)$ is the likelihood of that observation when the truth would be given by *x*.

The likelihood via observation space

Assume high-resolution observations to be assimilated in a (relatively) low-resolution model.



We have to calculate: $\boxed{p(y|\overline{y} = H(x))}$

so we need a relation between the two scales.

Relation between scales

Introduce the high-resolution variable in observation space

$$\boxed{z = H(x) + \tilde{z}}$$

Now use:

$$\boxed{p(y|x) = \int p(y|x,\tilde{z})p(\tilde{z}|x)\, d\tilde{z} = \int p(y|z)p(\tilde{z}|x)\, d\tilde{z}}$$

$p(y|z)$ is the instrument error

$p(\tilde{z}|x)$ is the representation error

$p(y|x)$ is the convolution of the two

Gaussian errors

For Gaussian instrument and representation errors we find:

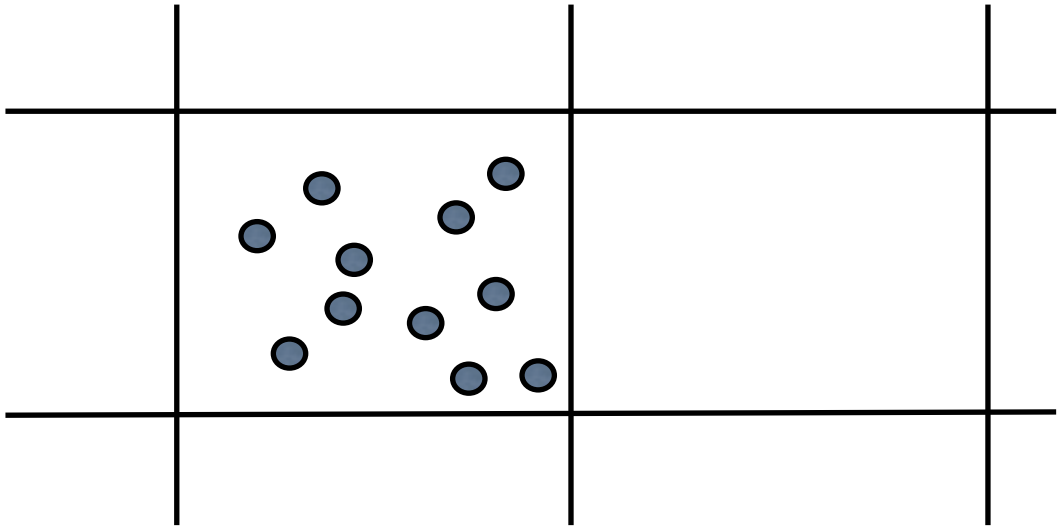
$$\boxed{p(y|x) \propto \int \exp\left(-\frac{1}{2}(y-z)^T R^{-1}(y-z) - \frac{1}{2}\tilde{z}^T C_o^{-1}\tilde{z}\right)\, d\tilde{z}}$$

Using $z = H(x) + \tilde{z}$ we can perform the integration:

$$\boxed{p(y|x) \propto \exp\left(-\frac{1}{2}(y-Hx)^T (R+C_o)^{-1}(y-Hx)\right)}$$

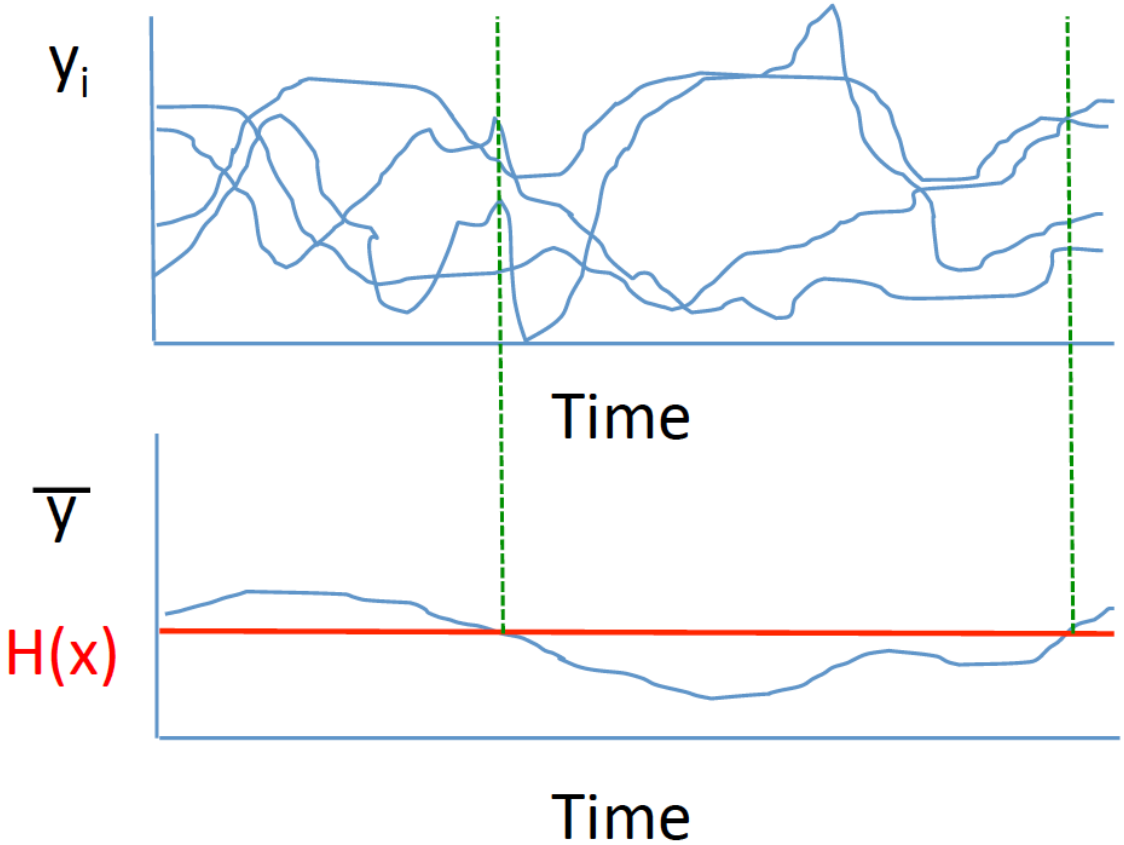
How can we calculate C_o ?

Use a whole set of observations:



And calculate their statistics (e.g. mean and variance)
Note that we need the statistics of *y* given the average over the model grid box: $p(y|\overline{y} = H(x))$

Calculation of C_0 from a time series



Only take those times into account when $\overline{y} = H(x)$
So at the times indicated by the green lines.

Warning:

Hence one should not take the variance of all observations but the variance of those observations that have an average equal to H(x)!

Superobbing

Can’t we just assimilate the average of the observations so do superobbing?

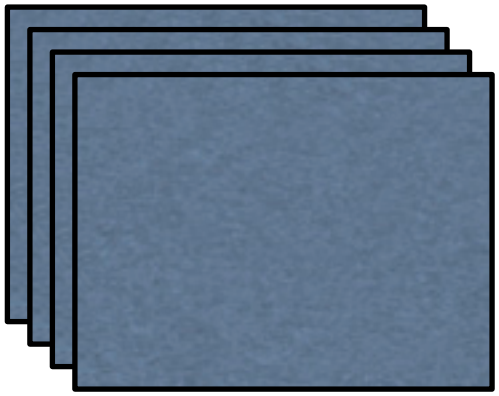
Yes , but the representation error is still there!!!

Naïve approach

Assume each measurement is a realisation of the box average, in which case

$$\boxed{\sigma_{\overline{y}}^2 = \frac{1}{M}\sigma_y^2}$$

However, we will get the same equation if we would use a set of realisations of the full model grid box, so this cannot be correct!



Variance of the mean

It is crucial to realise that we have to see each observation itself as a realisation of the model box average, so we have to condition this variance on the model state!

Assume that each of the observations has variance, conditoned on the model state, of:

$$\boxed{E\left[\left(y_i - H(x)\right)^2\right] = \sigma^2_{y|x} = \sigma^2_y + \sigma^2_R}$$

And the representation error is correlated with correlation coefficient ρ between any two observations.

Variance from superobbing

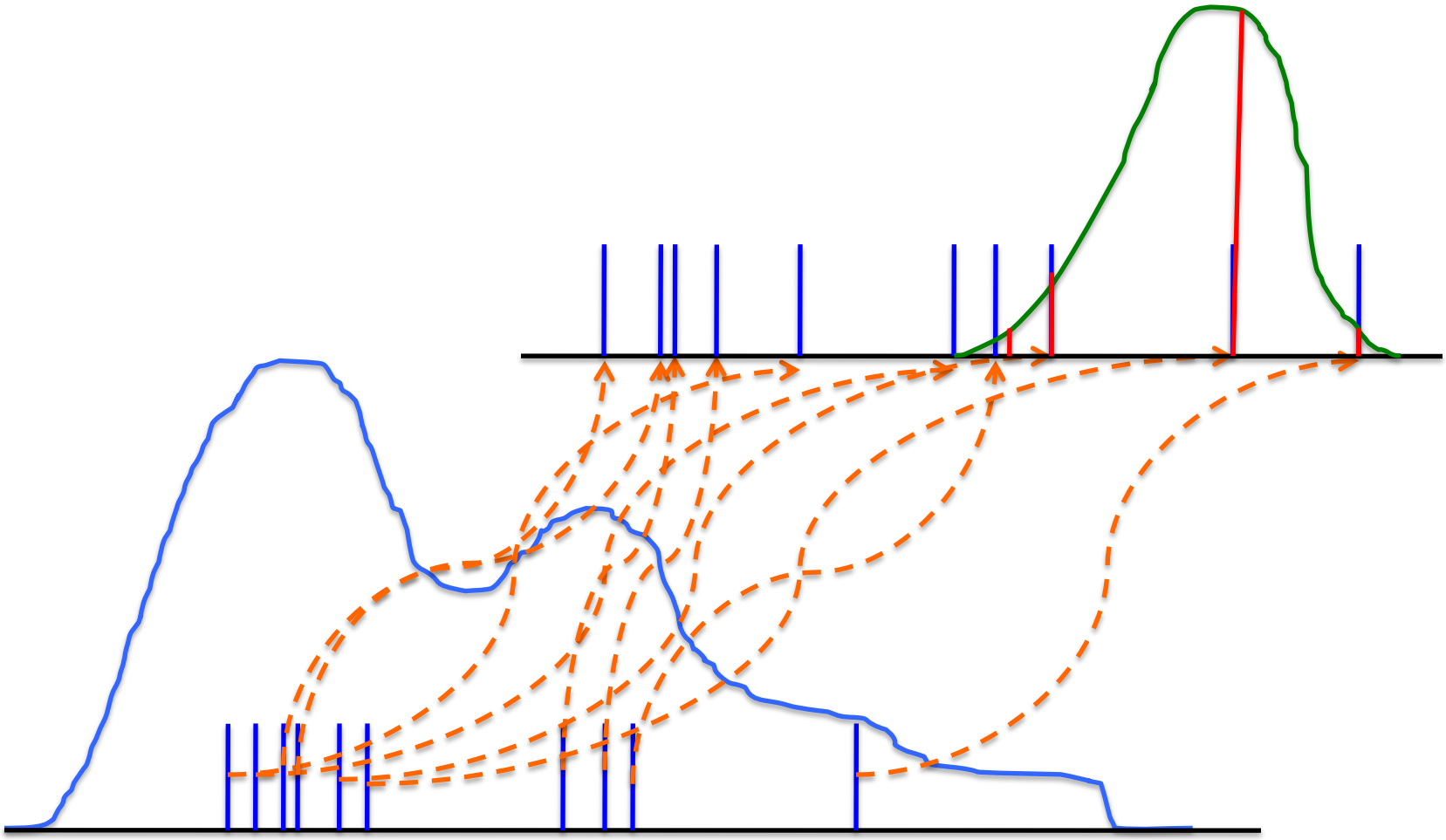
Hence we find:

$$\boxed{\sigma^2_{\overline{y}|x} = \frac{1}{M}\left(\sigma_y^2 + \sigma_R^2\right) + \frac{M-1}{M}\rho\sigma_R^2}$$

So superobbing is NOT free of representation errors. Note that if *M* increases the representation error remains and becomes the dominant term!

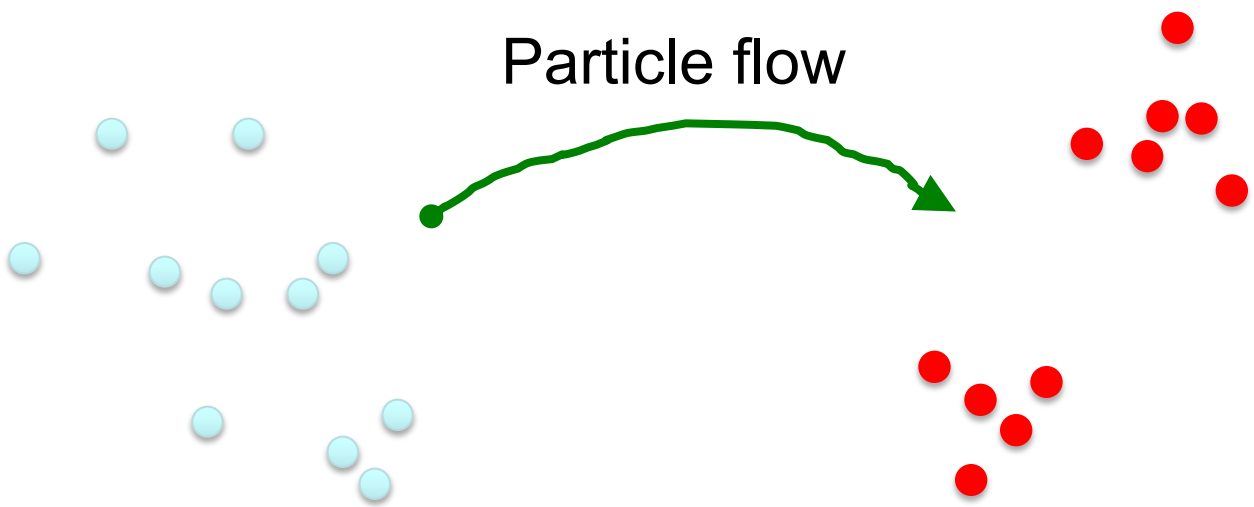
(Indeed assimilating the mean or all observations gives the same result as assimilation each separately.)

Nonlinear DA: Standard Particle filter



Ensemble size has to grow with $exp(N_y)$ to avoid degeneracy.

Transportation particle filters



Before observations
'Prior'

After observations
'Posterior'

Particle flows

The particles are propagated in artificial time *s* via

$$\frac{dx}{ds} = f_s(x)$$

The question is: How to choose the vector flow field $f_s(x)$?

The corresponding pdf of *x* evolves via the Liouville equation

$$\frac{\partial p_s(x)}{\partial s} = -\nabla_x \cdot (p_s(x) f_s(x))$$

with $p_0(x) = p(x)$ and $\lim_{s \to \infty} p_s(x) = p(x|y)$

Flow field should reduce *distance* between $p_s(x)$ and $p(x|y)$.

A ‘distance’ measure: KL divergence

The KL divergence between the present pdf and the posterior is given by:

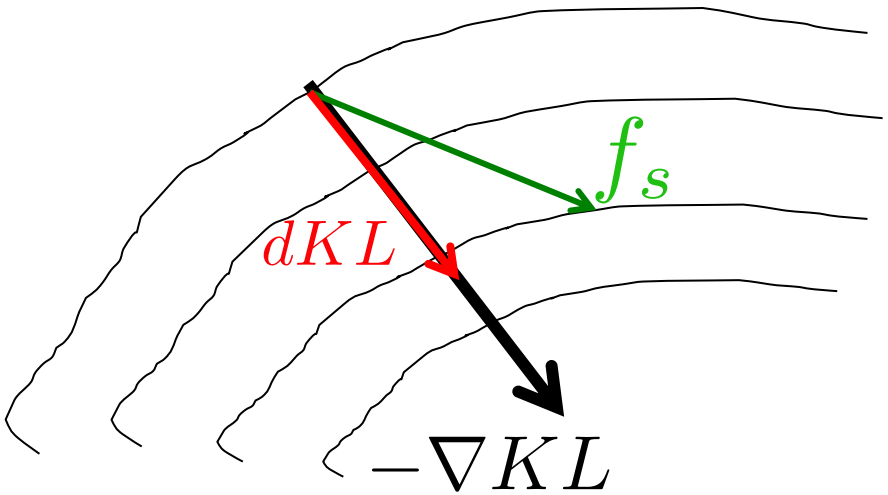
$$KL(p_s) = \int p_s(x) \log \left(\frac{p_s(x)}{p(x|y)} \right) dx$$

The change in KL due to the flow field $f_s(x)$ is given by

$$dKL(f_s) = \lim_{\epsilon \to \infty} \frac{KL(p_{s+\epsilon}) - KL(p_s)}{\epsilon}$$
$$= - \int p_s(x) \left[\left(\nabla_x \log p(x|y) \right)^T f_s(x) + \nabla_x f_s(x) \right] dx$$

We need to find the flow field that maximizes this change:
Complex optimization problem for the flow field…

The change in KL due to the flow field



$$dKL = \langle \nabla KL, f_s \rangle$$

Kernel embedding

Recall that the change in KL is given by:

$$dKL(x) = -\int p_s(x) \left[\left(\nabla_x \log p(x|y)\right)^T f_s(x) + \nabla_x f_s(x) \right] \, dx$$

And we have $dKL = \langle \nabla KL, f_s \rangle$.
Instead of solving the optimization problem directly we embed the flow in a Reproducing Kernel Hilbert Space, so

$$f_s(x) = \langle K(x, \cdot), f_s(\cdot) \rangle_{\mathcal{F}}$$

Using this kernel embedding we find:

$$\nabla KL = -\int p_s(x) \left[K(x', x) \nabla_x \log p(x'|y) + \nabla_x K(x', x) \right] \, dx'$$

Particle implementation

The evolution equation *for each particle* is now

$$\frac{dx_i}{ds} = f_s(x_i) = -\nabla KL(x_i)$$

This leads to an iterative scheme

$$x_i^{(j)} = x_i^{(j-1)} - \epsilon \nabla KL\left(x_i^{(j-1)}\right)$$

until $$|\nabla KL(x)| < \epsilon_{KL}$$

Convergence

LU, Lu and Nolan (2018) proof that:

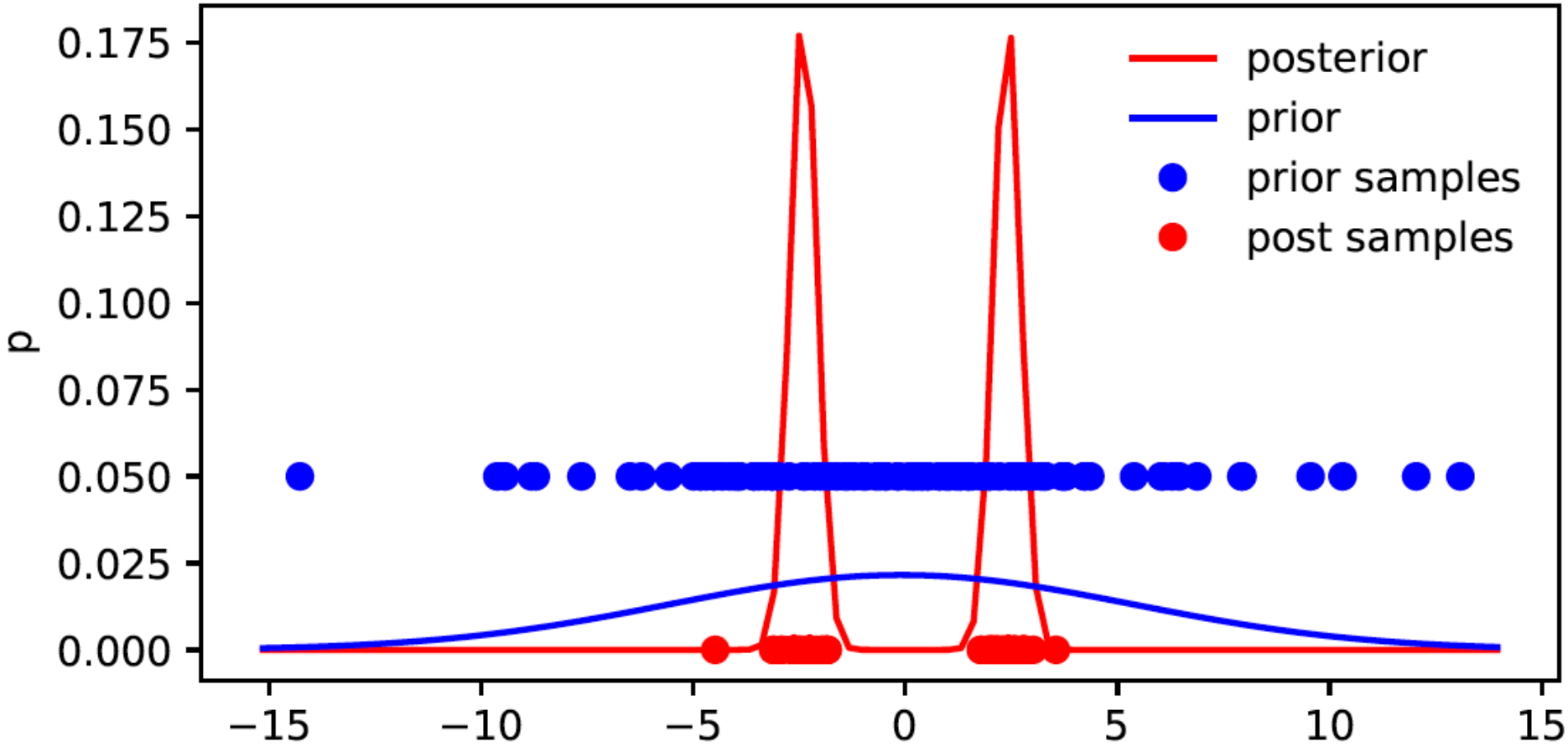
1. The large *N* limit converges to

$$\partial_t p_s = \nabla \cdot \left[p_s \left(\nabla \left\langle K, p_s \right\rangle - \left\langle K, p_s \nabla \log p(x|y) \right\rangle \right) \right]$$

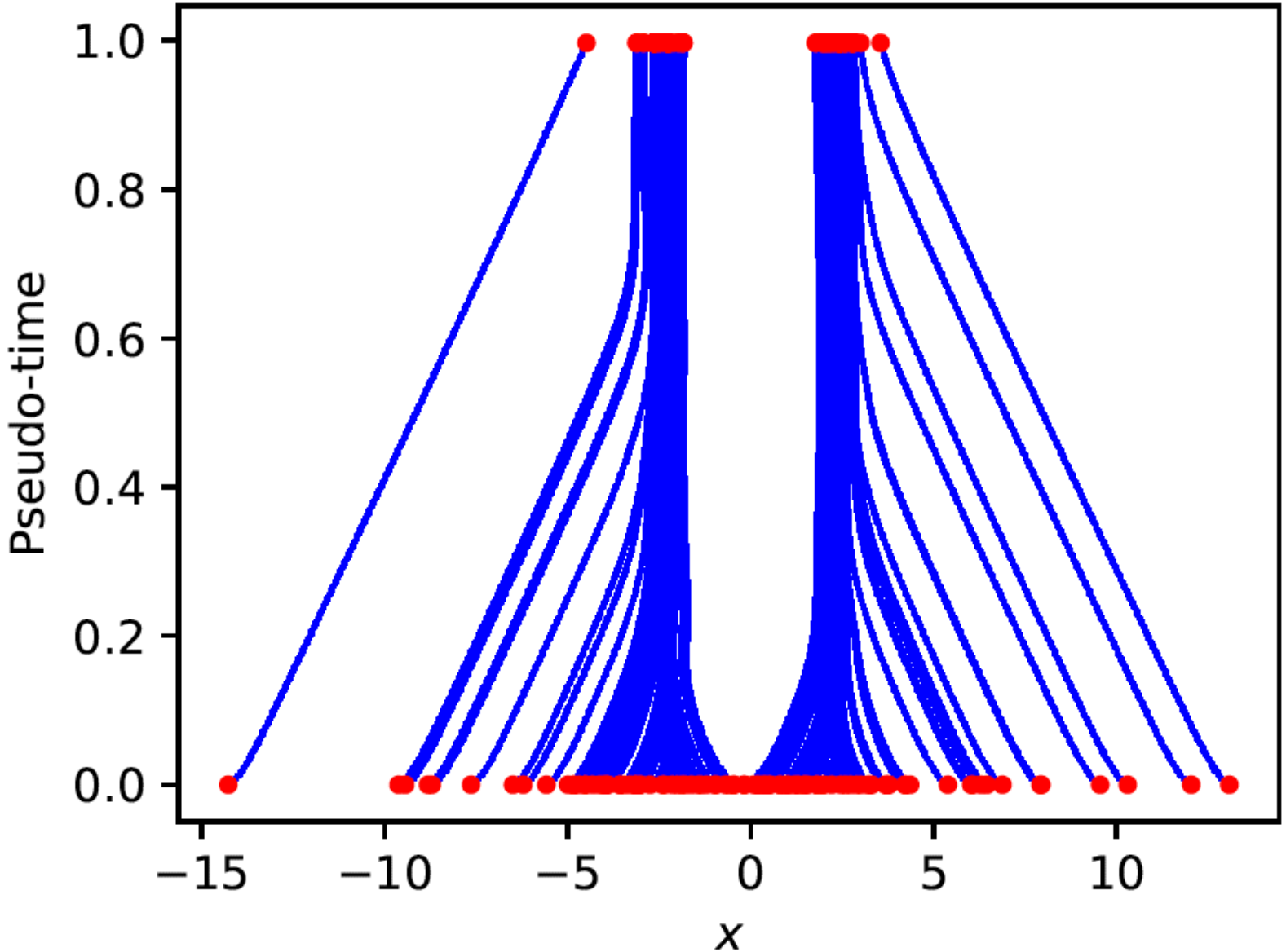
2. This PDE is well-posed and has unique solution

3. This solution converges to posterior pdf for large *s* limit.

Illustration 1-D case

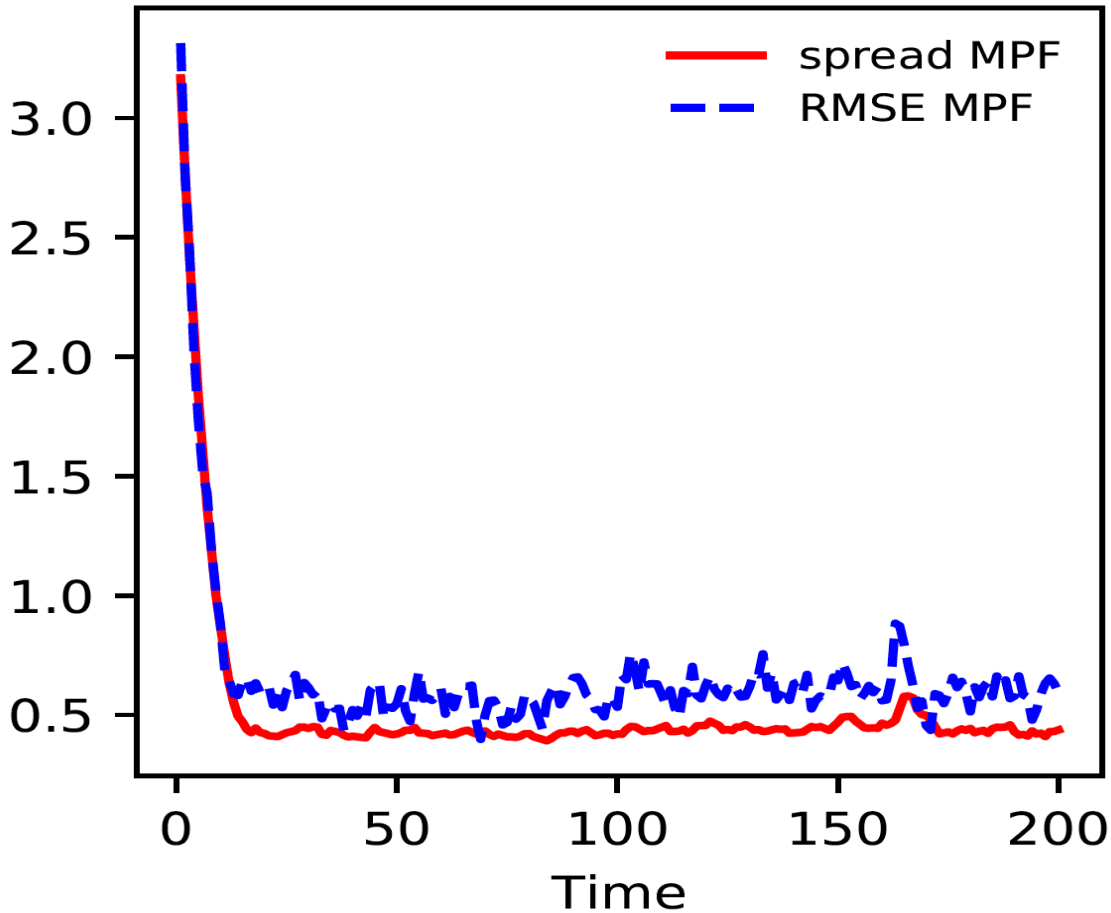


Particle Flow in 1-D case



Lorenz96 model 1000-dimensional N=20

All variables observed every ‘6 hour’.



Conclusions

- Great care must be taken in comparing observations and models.
- Data assimilation merges information from observations and models in most consistent way
- The ocean data-assimilation problem can be highly nonlinear.
- Particle Flows transport samples from the prior to samples from the posterior
- The iterative mapping PF is robust in low- to intermediate-dimensional systems. We are working on the real case!
- Cost equivalent to 3DVar on each particle.
- Model error covariance Q is essential.